# On Quantile Quantile plots for Generalized linear models

Nicole H. Augustin[1], Erik-André[2] Sauleau and Simon, N. Wood[1]

[1] Mathematical Sciences, University of Bath, Bath BA2 7AY U.K.

[2]Department of Biostatistics, Faculty of Medicine, University of Strasbourg, France

n.h.augustin@bath.ac.uk

September 14, 2011

**Abstract**

The distributional assumption for a generalized linear model is often checked by plotting the ordered deviance residuals against the quantiles of a standard normal distribution. Such plots can be difficult to interpret, because even when the model is correct, the plot often deviates substantially from a straight line. To rectify this problem García Ben and Yohai (2004, J. Comput. Graph. Stat. 13: 36-47) proposed plotting the deviance residuals against their theoretical quantiles, under the assumption that the model is correct. Such plots are closer to a straight line, when the model is correct, making them much more useful for model checking. However the quantile computation proposed in García Ben and Yohai is, in general, relatively complicated to implement and computationally expensive, so that general purpose software for these plots is only available for the Poisson and binary cases in the R package `robust`. As an alternative the theoretical quantiles can efficiently and simply be estimated by repeatedly simulating new response data from the fitted model and computing the corresponding residuals. This method also provides reference bands for judging the significance of departures of QQ-plots from ideal straight line form. A second alternative is to estimate the quantiles using quantiles of the response variable distribution

according to the estimated model. This latter alternative generally has lower computational cost than the first, but does not yield QQ-plot reference bands. In simulations the quantiles produced by the new methods give results indistinguishable from the original García Ben and Yohai quantile computations, but the scaling of computational cost with sample size is much improved so that a 500 fold reduction in computation time was observed at sample size 50000. Application of the methods to generalized linear models fitted to prostate cancer incidence data suggest that they are particularly useful in large dataset cases that might otherwise be incorrectly viewed as zero-inflated. The new approaches are simple enough to implement for any exponential family distribution and for several alternative types of residual, and this has been done for all the families available for use with generalized linear models in the basic distribution of R.

**Keywords:** Model checking, residuals, GLM.

# 1 Introduction

Consider a Generalized Linear Model (GLM) for $n$ response variable observations $y_i$, each with expectation $\mu_i$,

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}, \quad y_i \sim \mathrm{EF}(\mu_i, \phi),$$

where $\mathbf{X}_i$ is the ith row of a model matrix, dependent on known covariates; $\boldsymbol{\beta}$ is a vector of coefficients to be estimated; $\phi$ is a scale parameter; and $\mathrm{EF}(\mu_i, \phi)$ is some exponential family distribution dependent on $\mu_i$ and a known or unknown scale parameter $\phi$. $\boldsymbol{\beta}$ is estimated by maximum likelihood or maximum penalized likelihood estimation (for example if the model is a generalized additive model, or if some elements of $\boldsymbol{\beta}$ are to be treated as random effects), while $\phi$ can be estimated independently, typically using estimates based on either the model deviance or the Pearson statistic.

After estimation, all the information available for model checking is contained in the residuals (although there is little of it in residuals for a binary response; e.g. Cox and Snell (1989, page 73). The *raw* residuals are $r_i = y_i - \hat{\mu}_i$, where $\hat{\mu}_i$ is the model estimate of $\mu_i$. Because the distribution of these depends in a complicated way on the fitted model they are difficult to use for model checking unless the response is Gaussian. Therefore it is usual to standardize the

residuals, so that they will have constant variance and near constant distribution, if the model is correct. Two common standardizations are those used to produce *Pearson* and *deviance* residuals.

Pearson residuals utilize the fact that for any exponential family distribution, there exists a known function, $V$ such that $\mathrm{var}(y_i) = V(\mu_i)\phi$. In consequence the Pearson residuals,

$$p_i = (y_i - \hat{\mu}_i)/\sqrt{V(\hat{\mu}_i)},$$

will have constant variance if the model is correct.

Now consider deviance residuals. The model deviance, $D$, is twice the difference between the saturated log likelihood for the model and the maximized model log likelihood, all divided by the scale parameter (the saturated log likelihood is the maximized log likelihood for a model with one parameter per datum). For exponential family distributions it is always possible to write $D = \sum_i D_i$ where $D_i$ is a function of $y_i$ and $\hat{\mu}_i$ only. $D$ is constructed to behave rather like the residual sum of squares of a linear model, and by extension it is natural to view the quantities

$$d_i = \mathrm{sign}(y_i - \hat{\mu}_i)\sqrt{D_i}$$

as residuals. When the model is correct, the deviance residuals, $d_i$, have constant variance, and often have a distribution that is close to normal. The latter fact prompts the plotting of sorted deviance residuals against the quantiles of a standard normal, for model checking purposes. However, there are many applications of GLMs for which such plots show substantial deviation from a straight line, even when the model is correct (e.g. García Ben and Yohai, 2004). Modelling a response consisting of low counts is the most obvious example.

García Ben and Yohai (2004) propose avoiding the problems with normal QQ plots of the deviance residuals, by computing the empirical cumulative distribution function, $\hat{F}_D$ of the deviance residuals, conditional on the fitted model. They then generate $n$ quantiles $d_i^* = \hat{F}_D^{-1}((i - 0.5)/n)$ against which the sorted deviance residuals $d_i$ should be plotted. This should always yield a plot that is 'close' to a straight line, if the model is correct.

The García Ben and Yohai (2004) are never worse than normal QQ plots of the deviance residuals, and offer a substantial improvement in situations in which normal QQ plots are curved even when the model is correct. However the García Ben and Yohai method could usefully be

improved in two respects. Firstly, the method by which they compute the quantiles is moderately complicated to implement, and is relatively computationally expensive for a checking method. Specifically, in general, each evaluation of $\hat{F}_D$ requires $n$ evaluations of the quantile function and cumulative distribution function for the exponential family used in the model. i.e. each evaluation has $O(n)$ computational cost. In the absence of analytic shortcuts, accurate computation of the $d_i^*$ will require tabulating $\hat{F}_D$ at $O(n)$ points. Hence in general the computational cost of the $d_i^*$ is $O(n^2)$. Only the Poisson and binary cases seem to have been implemented in the R package robust (Wang, *et al*, 2010), and it is a relatively daunting task to implement all the other distributions routinely used with GLMs.

The second issue with the García Ben and Yohai (2004) QQ plots is that for count data there can be substantial *random* deviations from the ideal straight line, corresponding to discrepancies between the observed and expected number of observations of each count. This is easiest to see for binary data, where any deviation between the number of 1s observed and expected will cause some positive residuals to be assigned to negative quantiles, or vice-versa. Since these random discrepancies can sometimes be quite large, it would be useful for the plots to be accompanied by reference bands, indicating deviations that are larger than expected.

The remainder of this note shows how to approximate the $d_i^*$ simply in $O(kn\log(n))$ computer time and how to compute reference bands, where $k$ is a constant of order 10-100. The methods will be applicable to raw, Pearson or deviance residuals.

# 2 Obtaining quantiles

This section describes two alternative methods for generating quantiles for QQ plots. The first method requires only the ability to simulate new data from the fitted GLM, while the second also requires that the quantile function of the EF distribution is convenient enough to use. In this section the residuals are referred to as $d_i$, but the methods are general enough to employ with the Pearson or raw residuals also. Both methods are implemented in function `qq.gam` of R package `mgcv`.

## 2.1 Simulation based quantiles and reference bands

The first method is based on direct simulation. The idea is to directly simulate from $\hat{F}_D$, without forming it explicitly. $\hat{\boldsymbol{\mu}}$ and $\hat{\phi}$ are the estimates of $\boldsymbol{\mu}$ and $\phi$ from the original model fit.
For $j$ in 1 to $N_r$ repeat the following 2 steps.

1. For $i = 1, \ldots, n$ simulate new response data $\tilde{y}_i \sim \mathrm{EF}(\hat{\mu}_i, \hat{\phi})$.

2. Calculate the residual vector, $\tilde{\mathbf{d}}_j$ corresponding to $\tilde{\mathbf{y}}$, from $\tilde{\mathbf{y}}$, $\hat{\boldsymbol{\mu}}$ and $\hat{\phi}$.

For $i = 1, \ldots, n$ set $d_i^*$ to the $(i - 0.5)/n$ quantile of $\tilde{\mathbf{d}}^{\mathrm{T}} = (\tilde{\mathbf{d}}_1^{\mathrm{T}}, \tilde{\mathbf{d}}_2^{\mathrm{T}}, \ldots)$. The sorted $d_i$ are then plotted against the $d_i^*$, to yield the desired QQ-plot. Let $\hat{\mathbf{d}}_j$ denote the sorted version of $\tilde{\mathbf{d}}_j$. Reference bands for the QQ-plot can be obtained by plotting the $\hat{\mathbf{d}}_j$ vectors against $\mathbf{d}^*$ for $j = 1, \ldots, N_r$ (variant 1). Alternatively upper and lower quantiles only can be plotted, with the quantiles extracted from the sorted $\hat{\mathbf{d}}_j$ vectors in the obvious way (variant 2). The cost of the quantiles here will be $O(N_r n)$, where $N_r$ does not depend on $n$, so the cost is linear in the sample size. For both types of reference band there are $N_r$ sets of sorting to do, where the cost averages $O(n \log(n))$. In addition variant 2 requires $2n$ quantile computations, which in practice can be the dominant cost. Note that since it is only necessary to be able to simulate from the model for this method, it is very easy to implement for any exponential family.

## 2.2 Alternative computation of quantiles

If reference bands are not required then a more efficient alternative approach to estimation of the reference quantiles can be taken. To generate $n$ reference quantiles via the García Ben and Yohai method, $n$ quantiles of a uniform distribution $u_i$, can be generated in any order, and then used to obtain $d_i' = \hat{F}_D^{-1}(u_i)$, where $\hat{F}_D$ is the estimated cumulative distribution function for the residuals (marginalized over $i$). The resulting $d_i'$ are sorted to obtain the reference quantiles. An appealing efficient alternative is to set $d_i' = \hat{F}_{D_i}^{-1}(u_i)$ where $\hat{F}_{D_i}$ is estimated CDF of the $i^{\mathrm{th}}$ deviance residual. This alternative is computationally efficient because it simply sets $d_i'$ to the residual corresponding to the $u_i$ quantile of $y_i$, under the fitted model.

If the distribution of $D_i$ did not depend on $i$ then these quantiles would be exactly those of García Ben and Yohai, irrespective of the ordering of the $u_i$. However, in reality $F_{D_i}$ usually

depends on $i$, to some extent, so it is advisable to average the quantile estimates over several random permutations of the $u_i$, resulting in the following method.

Let $q_{\rm EF}(p, \mu, \phi)$ denote the quantile function of the exponential family used in the model (so that $\Pr\{y_i < q_{\rm EF}(p, \mu_i, \phi)\} = p$).

1. Set $u_i = (i - 0.5)/n$, for $i = 1 \ldots n$.
2. Repeat steps 3-5 for $j = 1 \ldots N_s$:
3. Randomly re-shuffle the $u_i$.
4. For $i = 1, \ldots, n$ set $\tilde{y}_i = q_{\rm EF}(u_i, \hat{\mu}_i, \hat{\phi})$.
5. Compute residuals $\tilde{d}_{ij}$ from $\tilde{y}_i$, $\hat{\mu}_i$ and $\hat{\phi}$, sorting the set $\{\tilde{d}_{ij} : i = 1 \ldots n\}$ into order.
6. The reference quantiles are now $d^*_i = \sum_{j=1}^{N_s} \tilde{d}_{ij}/N_s$, or are obtained as observed quantiles of the the complete set of $\tilde{d}_{ij}$.

Computational cost here is dominated either by the $O(N_s n \log(n))$ of sorting, or the $O(N_s n)$ response quantile evaluations. The direct simulation method of the previous section can be viewed as a more variable version of this method, in which the $u_i$ are replaced by $U(0, 1)$ random deviates. Notice also that if $F_{D_i}$ is the same for all $i$ then the quantiles estimated by this method have zero variance (and $N_s = 1$ could have been used).

# 3 Simulation comparison with García Ben and Yohai plots

The approach was briefly compared to the García Ben and Yohai (2004) method as implemented in function `qqplot.glmRob` of R package `robust` (Wang et al. 2010). Data were simulated independently from $y_i \sim$ binom$(\mu_i, n_i)$ where $i = 1 \ldots N$ and for each $i$, $n_i$ was randomly chosen to be 1, 2 or 3 with equal probability.

$$\text{logit}(\mu_i) = f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i})$$

where the $x_{ji}$ were i.i.d. $U(0, 1)$. The $f_j$ are shown in figure 1a-c. The generalized additive model $y_i \sim$ Poi$(\mu_i)$, where

$$\log(\mu_i) = f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + f_4(x_{4i}),$$

was fit to each replicate, using the R package `mgcv` (Wood, 2006), and the deviance residuals extracted. This setup was chosen because the fitted model mis-specification is not detectable from plots of residuals versus fitted values. Reference quantiles were computed by each of the 3 alternative methods (using $N_r = 100$, for direct simulation and $N_s = 10$ for the alternative). Section 2 reference quantiles were compared to the García Ben and Yohai quantiles using the p-value of a two sample Kolmogorov-Smirnov test as the metric.

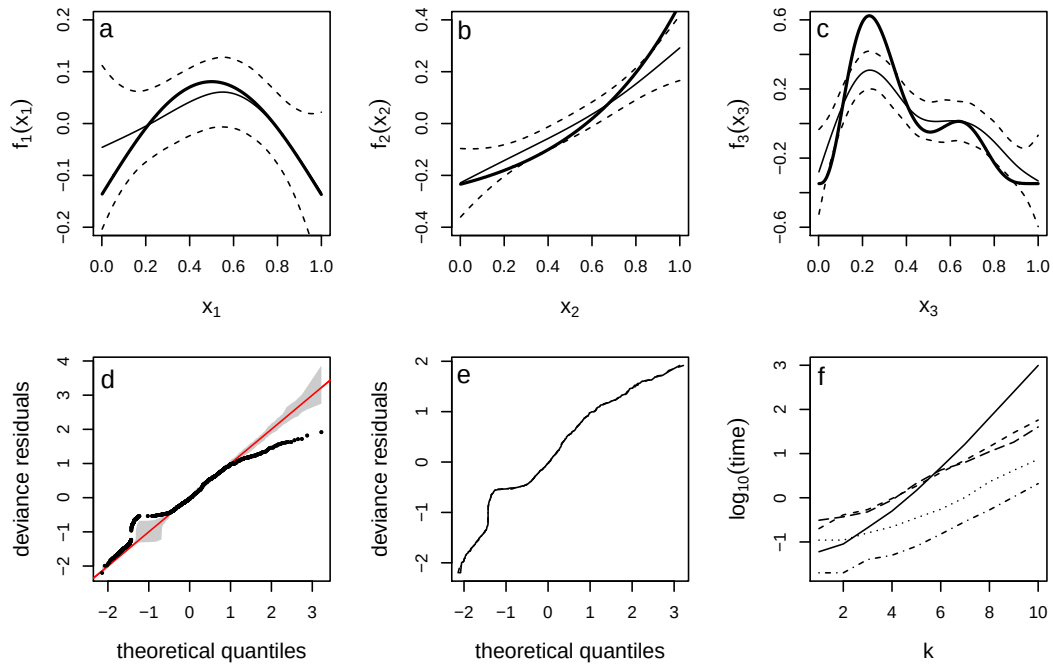


Figure 1: **a-c**: bold curves are the true $f_1$ to $f_3$ used in the section 3 simulations. The thin curves with dashed confidence limits show typical reconstructions for sample size 1000 using the wrong model as in section 3. **d**: Typical section 3 replicate QQ-plot with reference bands, produced by the method of section 2.1, showing that the plot detects the model mis-specification. **e**: QQ-plots produced by the methods of García Ben and Yohai (2004), section 2.1 and section 2.2, plotted with different line styles: they are practically indistinguishable. **f**: Plots of $\log_{10}$ computation time against sample size $N = 100 \times 2^k$. In ascending order at the right hand end of the plot: dash-dot is the section 2.2 method; dotted is the section 2.1 method, variant 1; long dashed is the time taken to fit the model itself; short dashed is the section 2.1 method, variant 2; continuous is the García Ben and Yohai (2004) method from the `robust` package.

For sample sizes $N = 100$, $400$ and $1000$ this experiment was repeated for 100 replicates. The methods produced such similar quantiles that the Kolmogorov-Smirnov p-value was 1 (to one part in $10^7$) for all 600 comparisons. Figure 1d shows a QQ plot with 90% reference bands for a typical replicate ($N = 1000$), illustrating that the plots can detect the model mis-specification. Figure 1e compares the QQ plots produced by the 3 methods for the same

replicate. They are practically indistinguishable. This is typical, although for some replicates the plots are just distinguishable in the extreme tails. In contrast the computational times are very different between the methods. 11 further replicates were run with $N = 100 \times 2^k$ for $k = 0 \ldots 10$, and the execution time for each method was recorded (the Kolmogorov-Smirnov p-value was again used to measure similarity of the quantiles estimated by the alternative methods and again the p-value was 1 for all replicates). The timing results are plotted in Figure 1f. For large sample sizes, the original García Ben and Yohai quantiles cost substantially more computer time than model fitting. Note that the cost of the section 2.1 variant 2 method was dominated by the cost of evaluating empirical quantiles, so timings for the more efficient variant 1 are also included (with $N_r = 50$).

These simulations suggest that the section 2.2 method is much more efficient than the original García Ben and Yohai method, at no detectable statistical performance cost. At $N = 50000$ the García Ben and Yohai method required over 1000 seconds of computer time, compared to 2 seconds for the section 2.2 method. The section 2.1 methods are also much more efficient than García Ben and Yohai for large data sets, and are at worst of comparable cost to model fitting: they offer the substantial advantage of also computing reference bands for the QQ-plots.

# 4 Example

The proposed QQ-plots were applied to deviance residuals of a generalised linear model fitted to prostate cancer incidence. The data were collected by the Cancer Registry of Haut-Rhin, France. This Registry covers the population of a region in the north-east of France. Prostate cancer is the most common of all cancers in France. Its incidence has increased by 8.5% between 2000 and 2005 and mortality decreased by 2.5%, in particular due to the introduction of screening. Screening is used to detect subclinical cancers but also generates over-diagnosis (artificially increases the incidence) and eventually over-treatment. The dataset contains all cases of prostate cancer (C61 in the ICD-10 classification) diagnosed between 1st January 1988 and 31st December 2005. Counts of prostate cancer are available by age, year of diagnosis and geographical unit. Age is categorized into 18, 5-year intervals, up to 85 years or more.

Let $O$ be the number of observed cases and E the number of expected cases, and identify

each of these by $O_{atr}$ and $\mathrm{E}_{atr}$ with cases indexed by the covariates age category $a$ (1 up to $A = 18$), year of diagnosis $t$ (1 up to $T = 18$) and geographical unit $r$ (from 1 to $R = 377$). Population counts by age and geographical unit are known for censuses and interpolated and extrapolated for other years. Letting $\mathrm{N}_{atr}$ denote the population counts, the corresponding expected counts are obtained as $\mathrm{E}_{atr} = \hat{p}\mathrm{N}_{atr}$ where $\hat{p}$ is the internally estimated global risk: $\hat{p} = \sum_{a=1}\sum_{t=1}\sum_{r=1} O_{atr} / \sum\sum\sum \mathrm{N}_{atr}$. The total number of incidences is 6,901 and the population at risk during the 18 years is estimated at 6,169,586. By geographical unit, this population varies from about 22 to about 54,109 by year. Due to covariates, the data set counts are spread over 122,148 cells.

Exploring the marginal distribution of the standardised incidence ratios (SIRs) shows that there are some trends in age category, time and space. Figure 2 shows the SIR at geographical unit $r$, for example the SIR at geographical unit $r$ is: $SIR_r = \sum_{a=0}\sum_t O_{atr} / \sum\sum \mathrm{E}_{atr}$ with 95% confidence intervals estimated according to Breslow and Day (1987). The geographic distribution of the SIRs and their confidence intervals shows that there may be some East - West trend, with higher SIRs in the West, and a number of communes exhibiting a SIR significantly above 1. Marginal plots of SIRS in time show that the SIRs increase in time, with a marked increase after 2000. Investigating the SIRs by age category shows that there is very little or no prostate cancer observed in age categories below 45 years and from 45 years there is an increase in the SIRs. Hence the data were aggregated for all age categories below 45 years into one category. This still leaves an extremely sparse data set with 96% of zeros in a total of 67,860 cells.

The observed counts can be represented with a multiplicative model (see for example Lawson, 2009)

$$\mathbb{E}[O_{atr}] = \mu_{atr} = \mathrm{E}_{atr} \times SIR_{atr},$$

leading to the GLM

$$\log(\mu_{atr}) = \log(\mathrm{E}_{atr}) + x_{atr}^T \boldsymbol{\beta}, \tag{1}$$

where $\log(\mathrm{E}_{atr})$ is assumed to be a constant and fitted as an offset (see above). The observed incidences $O_{atr}$ are assumed to be independently distributed and to follow a Poisson distribution

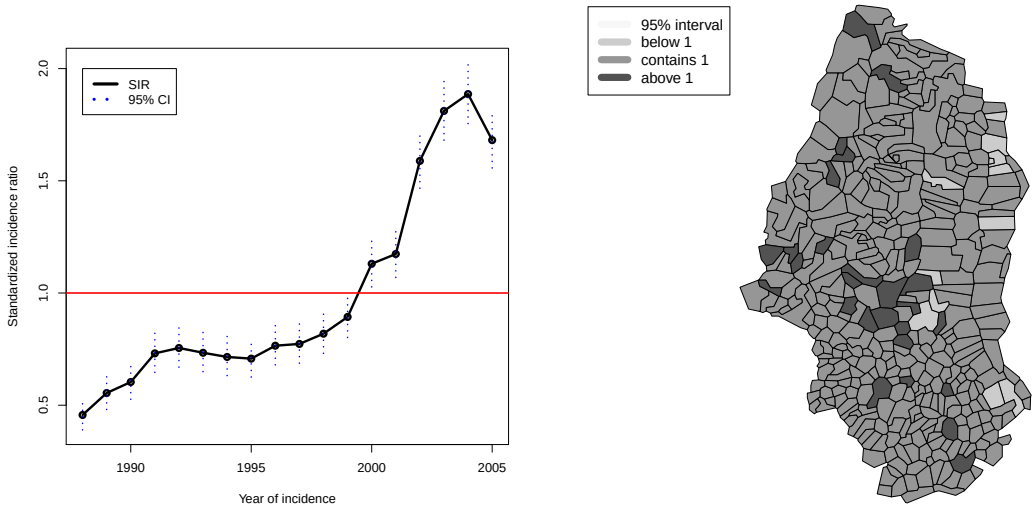


Figure 2: Standardized incidence ratio by year (left) and geographical unit (right). In the right plot discretisation is according to whether the 95% confidence interval of the $SIR_r$ is entirely above 1, containing 1 or entirely below 1.

with mean $\mu_{atr}$. Exploratory analysis suggests that the following terms could plausibly be included in the model: $x_{atr}$ contains longitude, age and year with polynomials of age and year up to cubic and single order interaction between age and year. Model 1 is compared with two alternatives: model (2) where longitude is dropped and model (3) where all terms related to year are omitted. The dispersion parameter estimated by quasi-likelihood for model 1 is close to one indicating that the Poisson distribution is adequate. Investigating the normal QQ-plots in the top of Figure 3 for these models shows that due to the extreme sparsity of these data the residuals do not follow a normal distribution and these standard residuals plots are very difficult to interpret. In comparison the proposed QQ plots in Figure 3 (bottom) give a clearer picture. They show that model 3 clearly does not fit. The difference between model 1 and 2 is marginal. Model inference confirms this as well: Dropping longitude (model 2) is just significant with a p-value 0.05 from a $\chi^2$-test and dropping all terms related to year (model 3) has a big effect with a p-value smaller than 2.2e-16. Notice that, without the lower plots, there would be a danger of incorrectly concluding that non of the models fit, and that a zero inflated distribution is needed in place of the Poisson.

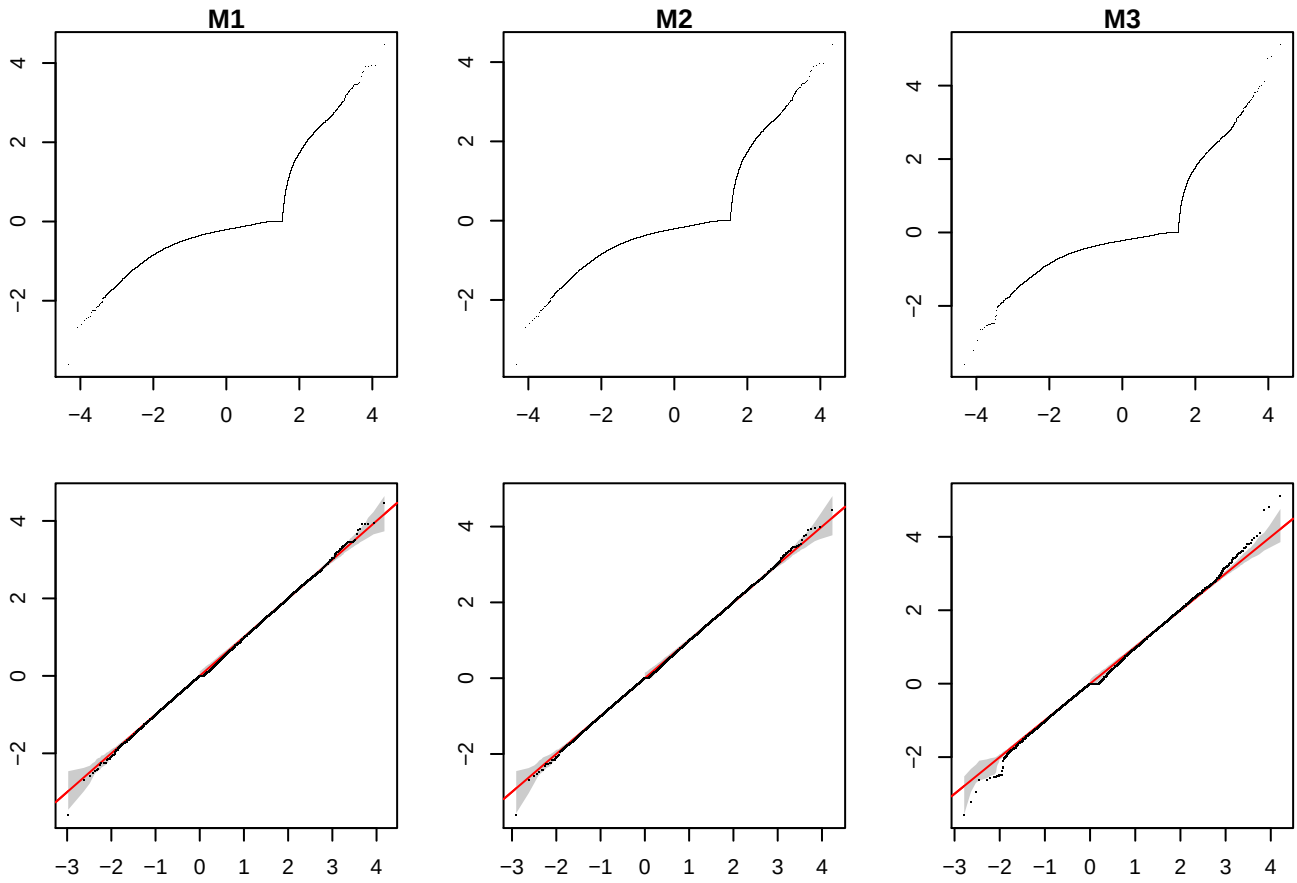


Figure 3: Residual models for model 1 to 3 (M1, M2, M3). Top: Plot of normal QQ plots of deviance residuals. Bottom: New proposed QQ plots of deviance residuals against their simulated quantiles assuming the model is correct with 90% reference band as described in section 2.1. Section 2.2 plots are graphically indistinguishable (but lack reference bands).

## References

Breslow, N.E. and N.E. Day (1987). *Statistical Methods in Cancer Research. Vol. II, The Design and Analysis of Cohort Studies (IARC Scientific Publication No. 82)*. Lyon: International Agency for Research on Cancer.

Cox, D.R. and E.J. Snell (1989) *Analysis of Binary Data*. Second Edition. London: Chapman & Hall.

Wang, J. and R. Zamar and A. Marazzi and V. Yohai and M. Salibian-Barrera and R. Maronna

and E. Zivot and D. Rocke and D. Martin and M. Maechler and K. Konis.(2010) *Robust: Insightful Robust Library. R package version 0.3-11.* url = http://CRAN.R-project.org/package=robust.

García Ben, M. and V.J. Yohai (2004) Quantile-Quantile Plot for Deviance Residuals in the Generalized Linear Model. *Journal of Computational and Graphical Statistics* 13(1):36-47

Lawson, A.B. *Bayesian Disease Mapping. Hierarchical Modeling in Spatial Epidemiology*. Boca Raton: Chapman & Hall/CRC

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R. Boca Raton: CRC/Chapman & Hall.